"Express Mail" Mailing Label No. **EL960827969US**

PATENT APPLICATION
ATTORNEY DOCKET NO. SUN-P9247-SPL

METHOD AND APPARATUS FOR ROUTING DATA ACROSS AN N-DIMENSIONAL GRID NETWORK

Inventor: Bernard Tourancheau

GOVERNMENT LICENSE RIGHTS

[0001] This invention was made with United States Government support under Contract No. NBCH020055 awarded by the Defense Advanced Research Projects Agency. The United States Government has certain rights in the invention.

BACKGROUND

Field of the Invention

[0002] The present invention relates to mechanisms for transferring data within a computing system. More specifically, the present invention relates to a method and an apparatus for routing data across an n-dimensional grid network.

Related Art

[0003] Dramatic increases in microprocessor clock speeds have not been matched by corresponding increases in chip-to-chip communication speeds. Consequently, inter-chip communication delays are rapidly becoming a major bottleneck to overall computer system performance. For example, it is now common for accesses to off-chip memory to require hundreds of processor clock cycles. This means that microprocessors are often stalled waiting for memory operations to complete.

[0004] The performance-limiting effects of these long inter-chip communications delays can be somewhat mitigated by providing large on-chip caches, and by providing mechanisms to support out-of-order execution, so that useful work can be accomplished during accesses to off-chip memory. However, as memory accesses begin to take hundreds of processor cycles to complete, even these large on-chip caches and out-of-order execution cannot keep a processor busy on typical processor workloads.

[0005] System developers are beginning to consider different interconnection topologies to provide fast chip-to-chip communication within computer systems. One promising topology is a two-dimensional grid network, wherein chips are coupled to four of their nearest neighbors (North, East, South, and West). The close proximity between chips in a two-dimensional grid facilitates high-throughput and low-latency communication between adjacent chips. Moreover, a two-dimensional grid network can be easily implemented with existing packaging technologies, which are well-suited to planar layouts. Unfortunately, existing routing mechanisms for two-dimensional grid networks can be quite complicated because they must deal with collisions, load-balancing issues and must avoid deadlock conditions.

[0006] Memory systems are a major limitation to performance in modern computer systems. Although the speed of processors has increased dramatically, the latency of memory access has not been reduced commensurately. The result is that computer processors spend relatively longer times waiting for a response from their memory systems. In a modern computing system, the processor may lose hundreds of potential machine cycles waiting for some piece of data from memory.

[0007] Hence, what is needed is a method and an apparatus that facilitates fast chip-to-chip communications within a computer system without the problems described above.

SUMMARY

[0008] One embodiment of the present invention provides a system for routing data between integrated circuit devices. This system couples together an n-dimensional grid of integrated circuit devices using multiple independent communication networks, wherein each of the communication networks moves data in only two orthogonal directions (e.g., when n = 2, the directions can be North and East, North and West, South and East, or South and West). The system also includes a routing mechanism that routes data across these communication networks, as well as, into, out of, and through integrated circuits within the n-dimensional grid of integrated circuits. Note that the process of routing a signal across a given network is greatly simplified because it is not possible to create a cycle that causes a deadlock within a given network.

[0009] In a variation of this embodiment, the n-dimensional grid of integrated circuit devices includes memory devices.

[0010] In a further variation, the n-dimensional grid of integrated circuit devices includes processor devices.

[0011] In a further variation, the multiple communication networks for a two-dimensional grid include a communication network configured to move signals North and East, a communication network configured to move signals North and West, a communication network configured to move signals South and East, and a communication network configured to move signals South and West.

[0012] In a further variation, the routing mechanism is configured to statically route data items across the multiple communication networks.

[0013] In a further variation, the routing mechanism is configured to dynamically route data items through network junctions within each integrated circuit.

[0014] In a further variation, a header attached to each data item indicates a number of horizontal steps and a number of vertical steps required for the data item to reach its destination. During the dynamic routing process, at each network junction, the routing mechanism removes a horizontal step or a vertical step from the header for the data item, depending on which dimension is dynamically selected.

BRIEF DESCRIPTION OF THE FIGURES

[0015] FIG. 1 illustrates circuit devices coupled together in accordance with an embodiment of the present invention.

[0016] FIG. 2 illustrates communication directions for the communication networks in accordance with an embodiment of the present invention.

[0017] FIG. 3 illustrates a single network implementation within a circuit device in accordance with an embodiment of the present invention.

[0018] FIG. 4 illustrates the networks within a circuit device in accordance with an embodiment of the present invention.

[0019] FIG. 5 illustrates a possible path between two circuit devices in accordance with an embodiment of the present invention.

[0020] FIG. 6 presents a flowchart illustrating the process of creating a routing header and sending data in accordance with an embodiment of the present invention.

[0021] FIG. 7 presents a flowchart illustrating the process of routing data within a circuit device in accordance with an embodiment of the present invention.

DETAILED DESCRIPTION

[0022] The following description is presented to enable any person skilled in the art to make and use the invention, and is provided in the context of a particular application and its requirements. Various modifications to the disclosed embodiments will be readily apparent to those skilled in the art, and the general principles defined herein may be applied to other embodiments and applications without departing from the spirit and scope of the present invention. Thus, the present invention is not intended to be limited to the embodiments shown, but is to be accorded the widest scope consistent with the principles and features disclosed herein.

[0023] The data structures and code described in this detailed description are typically stored on a computer readable storage medium, which may be any device or medium that can store code and/or data for use by a computer system. This includes, but is not limited to, magnetic and optical storage devices such as disk drives, magnetic tape, CDs (compact discs) and DVDs (digital versatile discs or digital video discs), and computer instruction signals embodied in a transmission medium (with or without a carrier wave upon which the signals are modulated). For example, the transmission medium may include a

communications network, such as the Internet, storage SA and NAS, LAN, system SAN, backplane, SOC, etc.

Circuit Devices

[0024] FIG. 1 illustrates a number of circuit devices coupled together in accordance with an embodiment of the present invention. Note that circuit devices 102, 104, 106, 108, 110, 112, 114, 116, and 118 are coupled together by multiple independent networks. These multiple, independent networks are described in detail below in conjunction with FIGs. 2-4. Note that the any number of circuit devices can be included to form a grid of circuit devices as large as desired. Note also that while the following description relates to a two-dimensional grid, it will be obvious to a person with ordinary skill in the art that the system can be extended to any number of orthogonal dimensions.

[0025] In one embodiment of the invention, circuit devices 102, 104, 106, 108, 110, 112, 114, 116, and 118 are memory devices that are driven from the edge of the grid by processors (not shown). In another embodiment of, one or more of the circuit devices, such as circuit device 110, is a processor.

[0026] The communication network illustrated in FIG. 1 has a number of advantages. In a two-dimensional grid, the average distance (hamming distance) between two circuit devices grows linearly, while the number of the circuit devices grows as the square of the distance. Hence, while the communication delay increases linearly, the number of devices that can be reached increases quadratically. Moreover, the close proximity of neighboring chips allows fast chip-to-chip communication. In order to improve performance, data that is accessed most often can be placed in a memory device that is near to the processor circuit device, whereas data that is rarely accessed can be place in more distant memory devices. While the description herein relates to a two-dimensional grid,

it should be noted that the system can easily be extended to an arbitrary n-dimensional grid with no changes to the basic details of the present invention.

Network Communication Directions

[0027] FIG. 2 illustrates the communication directions associated the different communication networks in accordance with an embodiment of the present invention. The system illustrated in FIG. 1 includes four networks 202, 204, 206, and 208. The communication directions for networks 202, 204, 206, and 208 are illustrated in FIG. 2.

[0028] Network 202 moves data only North or East. Network 204 moves data only North or West. Network 206 moves data only South or West. Finally, network 208 moves data only South or East. Note that since the distance from source to destination increases monotonically as the data is routed along the network, there can be no cycles in the graph and thus, no deadlocks are possible.

[0029] The available paths between peers in these networks are shortest paths relative to Hamming distance. Moreover, regarding the two-dimensional (respectively, n-dimensional) grid, all of the shortest paths are available.

Single Network

[0030] FIG. 3 illustrates a single network within a circuit device in accordance with an embodiment of the present invention. The network illustrated in FIG. 3 is network 208 from FIG. 2. Network 208 can move data only South and East. Each of networks 202, 204, 206 and 208 operates in a similar manner so only the operation of network 208 will be discussed in detail herein.

[0031] During operation of network 208, data enters circuit device 108 either from its Western neighbor or its Northern neighbor. Once data enters circuit device 108, the data can be moved into RAM 302, or can exit circuit device 108 either to the East or the South. Note that in some embodiments,

RAM 302 can be a processor or other type circuit device. Two direction-switching branches are included on network 208 within circuit device 108. One of these branches switches the data from the Western neighbor to the Southern neighbor, while the other branch switches the data from the Northern neighbor to the Eastern neighbor. Note that the data can pass through circuit device 108 without switching direction. Two I/O branches are also provided to move data into and out of the circuit (RAM 302 in this case).

[0032] When data needs to be routed between two circuit devices, the source calculates the horizontal distance and the vertical distance of the destination using Cartesian coordinates for each of the two circuit devices. The source then creates some information attached to the data, for instance a header, which includes the number of horizontal steps and the number of vertical steps between these circuit devices. Note that in one embodiment of the present invention the number of steps is represented by a single bit for each step. However, in an alternate embodiment, the number of steps can be encoded differently, for instance by an integer. Note also that only the absolute value of the number of steps is encoded in the header.

[0033] The sign of the number of horizontal steps and vertical steps is used to determine which network is used to transfer the data.

[0034] The routing into the circuit device can be implicit, i.e. when no more routing information is available, we route into the device itself. For instance, when data and its associated header enter a circuit device, the number of steps remaining is examined. If both the horizontal and vertical number of steps is zero—all of the bits are set to zero in the case where each step is represented by a bit—the I/O branch routes the data into the element within the circuit device. For example, if all of the bits are set to zero in the header associated with data entering circuit device 108, the data is routed to RAM 302. Otherwise, the data is

routed through the major branches based on the bits that are set and possible contention for an output path. This process is described in more detail in conjunction with FIG. 6 below.

Networks Within a Circuit Device

[0035] FIG. 4 illustrates multiple networks combined within a circuit device in accordance with an embodiment of the present invention. In particular, circuit device 108 includes RAM 302 and four networks. Each of the four networks operates essentially as described above for network 208. The input and output directions for each of the four networks is as described above in conjunction with FIG. 2. Note that there is no direct connection between any of the networks. Note also that the networks operate asynchronously and independently.

Paths Between Circuit Devices

[0036] FIG. 5 illustrates a possible path between two circuit devices in accordance with an embodiment of the present invention. A grid of circuit devices is coupled together by the four networks as describe above. In FIG. 5, a single line is shown between each circuit device in place of the multiple networks in order to simplify the diagram. Assume that data need to be passed from circuit device "X" to circuit device "Y." The source circuit device (X) calculates the distance between device "X" and device "Y" to be two horizontal steps and 6 vertical steps and encodes this data in the header. Since the destination is to the right and up from the source, the source selects network 1 to send the data (and the associated header).

[0037] As illustrated in FIG. 5, the data is moved to the circuit device to the right of the source. The bit representing this horizontal transition is cleared by

shifting the bit out of the header and shifting in a zero bit. Note that other methods of reducing the number of remaining steps can be used. As the data is routed between device "X" and device "Y," each circuit device forwards the data and the header in one of the two directions until the count in each direction is reduced to zero. This routing can depend on any of several strategies. For instance, the routing can be based on information from the data packet, the local switch and/or link state, or the global system state. The routing decisions that are made at each circuit device can thus range/include static source routing and dynamic reaction to contentions as in hot-potato routing. These are all shortest path.

[0038] When the count is reduced to zero in each direction, the data has reached destination "Y." Note that there are several paths that can be taken between "X" and "Y." The path chosen depends upon contention for the network paths and the routing decisions that are made at each circuit device.

Creating a Header

[0039] FIG. 6 presents a flowchart illustrating the process of creating a routing header and sending data in accordance with an embodiment of the present invention. The system starts when data is received which is to be routed to a circuit device within the grid (step 602). Next, the system calculates the horizontal and vertical distances to the destination circuit device (step 604). The horizontal and vertical distances can be calculated by taking the difference between the Cartesian coordinates of the destination and the source (Y_D-Y_S and X_D-X_S). The system then creates a header for the data, which includes the number of horizontal steps and vertical steps between the source and destination (step 606).

[0040] The system then selects an output network for the data depending on the signs of the horizontal and vertical distances to the destination (step 608). Finally, the system sends the data over the selected network and reduces the count in the appropriate direction by one (step 610). Note that the process continues as describe below in conjunction with FIG. 7 until the data reaches the destination.

Routing Data Within a Circuit Device

[0041] FIG. 7 presents a flowchart illustrating the process of routing data within a circuit device in accordance with an embodiment of the present invention. Note that this a representative strategy. There are numerous strategies that will accomplish the objective. The system starts when a circuit device receives data from an adjacent circuit device (step 702). Next, the system examines the header to determine whether the H and V bits are set (step 704). If neither the H nor the V bits are set, the system routes the data into the cell concluding the data transfer (step 708).

[0042] If both the H and V bits are set, the system determines if there is contention on both the H and V outputs (step 710). If so, the system returns to step 710 to wait for the contention to end on either direction. Otherwise, the system selects an output direction for the data (step 712). Note that if only one direction does not have contention, that is the chosen direction. However, if neither H nor V has contention, a direction is selected using a predetermined means, such as random or round robin.

[0043] If only H bits are set at step 704, the system determines if there is contention in the H direction (step 714). If so, the system returns to step 714 to wait for the contention to end. Likewise, if only V bits are set at step 704, the system determines if there is contention in the V direction (step 716). If so, the system returns to step 716 to wait for the contention to end.

[0044] After selecting a direction at step 712 or after there is no contention at steps 714 or 716, the system reduces the count in the selected direction (step 718). Note that this decrement can be done in several places after the branch switching decision is taken. Finally, the system sends the data in the selected direction (step 720).

[0045] The foregoing descriptions of embodiments of the present invention have been presented for purposes of illustration and description only. They are not intended to be exhaustive or to limit the present invention to the forms disclosed. Accordingly, many modifications and variations will be apparent to practitioners skilled in the art. Additionally, the above disclosure is not intended to limit the present invention. The scope of the present invention is defined by the appended claims.